Three test-score metrics that all states should report in the COVID-19-affected spring of 2021

Andrew Ho, Harvard Graduate School of Education

Draft Memo[1] – February 26, 2021 Current version here.

In this memo, I propose three metrics for state reporting of aggregate test scores in this COVID-19 affected school year. These metrics advance the goal of accurate score interpretations and fair trend comparisons among schools and districts. Without metrics like these, efforts to, "target resources and supports to the students with the greatest needs" (U.S. Department of Education, 2021, p. 1), by policymakers and the public will fail.

The first metric is a *match rate* that helps users flag whether the percentage of comparable test scores in a school or district is unusually low. The second metric is a *fair trend* that compares scores from this year to those of similar students from two years prior. The third metric is an *equity check* that indicates best-case academic disparities given potentially large percentages of students who do not have comparable test scores. Without explicit presentation of these or similar metrics in prominent positions in public score reports, users will falsely identify academically thriving schools as needing academic support, and users will falsely identify schools that need academic support as academically thriving.

On Monday, the Biden administration took a bend-but-not-break approach to flexibility for state educational testing in this COVID-19-affected school year (U.S. Department of Education, 2021), inviting states to request a waiver from accountability while maintaining testing requirements and public reporting requirements. Their stated purpose of testing is to "address the educational inequities that have been exacerbated by the pandemic" (p. 1). Addressing growing inequities requires accurate measurement of inequities, including accurate comparisons of inequities over time. This is challenging in any year. This year, it will be essential to distinguish changes in academic proficiency from substantial and variable changes in the population that has comparable test scores.

This memo focuses narrowly on grade-, school-, and district-level test-score metrics for state standardized tests for the purpose of large-scale monitoring and resource targeting. Student-level test score reports deserve similar care but are beyond the scope of this memo. I have argued elsewhere that educational test scores should serve a tertiary purpose among a system of multiple measures in this pandemic (Klugman & Ho, 2020). I argue here similarly that states should view their spring efforts as an *educational census* rather than an *educational assessment*.

I recommend that states and districts report three metrics that answer clear questions and address looming threats to the validity of aggregate score interpretations. Table 1 reviews the three metrics, the questions they answer, the problem they solve, and how they solve the problem. Further technical details follow. These metrics require states to have longitudinal data systems and stable testing systems since the 2016-2017 academic year. I briefly review solutions for states that lack these systems. These metrics are not a panacea. They cannot overcome a failure of year-to-year equating. They require scores to have the same meaning in terms of student knowledge, skills, and abilities, as prior years.

[1] I presented a version of this memo to the Technical Issues in Large Scale Assessment (TILSA) collaborative of the Council of Chief State School Officers on Tuesday, February 23, 2021.

Table 1. *Three test-score metrics for aggregate reporting in a COVID-affected school year. Assumes longitudinal data and stable testing programs. Technical details below.*

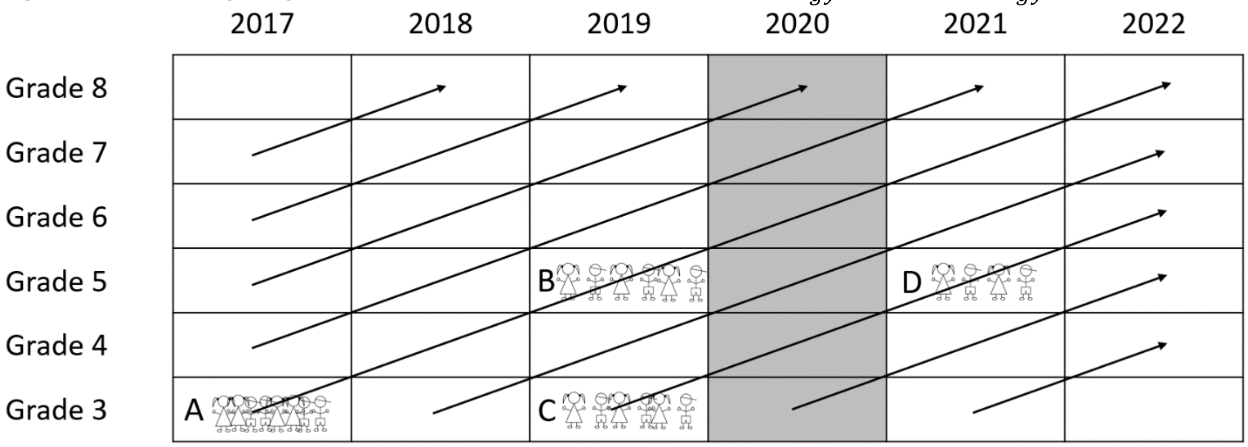
Metric	Question	Problem	Solution
1) Match Rate	What percentage of students have comparable test scores?	Students with comparable test scores are not representative of the usual tested population.	Prominently display the percentage of students with comparable test scores (and unusual declines in such percentages).
2) Fair Trend	How much academic progress have students in 2021 made compared to academic peers in 2019?	Comparing 2021 academic proficiency to existing 2019 baselines confuses changes in populations with changes in proficiency.	Report progress compared to fairer 2019 baseline proficiency rates comprised of academic peers.
3) Equity Check	What are best-case academic outcomes for students who do not have comparable scores?	We do not know the scores of *students who do not have comparable scores* . These may or may not be the most vulnerable students.	Report outcomes for academic peers of *students who do not have comparable scores* .

Metric 1: The Match Rate

Accurate interpretation of aggregate scores requires an accurate answer to the question, what population does this report describe? In a typical year, this answer is straightforward: the students in grade, school, or district.[2] The percentage of tested students is typically high, with targets above 95%. Longitudinally, many schools have nontrivial grade-to-grade mobility rates, where students leave or enter some schools more frequently between school years than other schools. However, mobility rates are typically steady over time.

In this school year, there may be substantial numbers of enrolled students who may not have comparable scores, whether they opt out of testing or test remotely in noncomparable conditions. In addition, there may be substantial and atypical numbers of students who are not in school who otherwise would be enrolled, above and beyond typical mobility rates. To illustrate this, Figure 1 shows the standard progression of cohorts across grades 3-8 over the past 5 school years, with academic-year-ending 2020 grayed out due to missed testing data.

Figure 1. *Illustrating longitudinal and cross-sectional match rates, m^*_{gy} (D to C) and m_{gy} (D to B).*



[2] From this point on, I will refer to school reports, although metrics generalize to grades, districts, and other levels of aggregation.

A naïve match rate for grade g and year y compares the number of students in that grade-year cell to the number from two years prior in the same grade:

$$m_{gy} = \frac{n_{gy}}{n_{g,y-2}}.$$

This match rate matches on grades, not individuals. It allows in-migration to offset out-migration of students, and it may exceed 100%. It may nonetheless be a useful benchmark when states do not have longitudinal data systems from early grades. Figure 1 shows the match rate for grade 5 in 2021 as the comparison of cell D to cell B: $m_{5,2021} = n_{5,2021}/n_{5,2019}$.

In contrast, the longitudinally linked match rate m^* is estimated in school s, grade g, and year y as:

$$m^*_{gy} = \frac{n^*_{gy}}{n_{g-2,y-2}}.$$

The variable $n_{g-2,y-2}$ is the number of students in each school in grade $g-2$, and year $y-2$, two years and two grades prior. Using longitudinal data, we can identify those who remain in grade g and y as n^*_{gy}. I use the * superscript to denote quantities that use longitudinal matching. Thus, the match rate m^* reports the proportion of students from two years prior who remain in grade g and year y. Reasons for low match rates include mobility and non-standard grade progression. Students who enter the school at any other year or grade are not counted in the match rate, such that the total number of students n_{sgy} is the sum of students n^*_{sgy} who were there two years and two grades ago, plus anyone else who may have entered that grade or year since, which I designate n'_{sgy}. I use the ' superscript to denote complementary quantities. Thus, the total number of students in the current year and grade is $n_{sgy} = n^*_{sgy} + n'_{sgy}$. Here, we assume that the percentage of students n'_{sgy}/n_{sgy} is low.

Figure 1 shows how we would construct the longitudinal match rate for grade 5 in 2021 as $m^*_{5,2021} = n^*_{5,2021}/n_{3,2019}$. First, we go back two years to grade 3 students in 2019 in the same school. Then, we find how many of them remain in grade 5 in 2021. Note that $m^*_{5,2021}$ matches cells D to C longitudinally in Figure 1, whereas $m_{5,2021}$ compares cells D to B, cross-sectionally.

Figure 1 also shows how we can construct a reference match rate from previous cohorts of data. This reference is important, because it prevents the false flagging of schools that have always had high mobility rates. This reference match rate is $m^*_{5,2019} = n^*_{5,2019}/n_{3,2017}$. Substantial declines from $m^*_{5,2019}$ to $m^*_{5,2021}$ could result in a flag for caution giving changing percentages, perhaps past some number of percentage points that states determine warrants caution.

For middle schools that may draw upon multiple elementary feeder schools, the sum of all $n_{g-2,y-2}$ across feeder schools may serve as the denominator. When elementary schools feed to many middle schools, comparing match rates in 2021 to reference match rates in 2019 will be particularly important. Where reference match rates are consistently low, including across school transitions, it may make sense to rescale the match rate to the baseline expectation set from two years prior, for all schools. This is interpretable as the percentage of the match rate from the baseline year. In idiosyncratic cases, this match rate may exceed 100%, due to unusual relative in-migration of students and/or small sample sizes. States can inspect these values and report them at a ceiling of 100%.

$$\widetilde{m}^* = \frac{m_{gy}^*}{m_{g,y-2}^*}.$$

For grades 3 and 4, where longitudinal data from two years and grades prior may not be available, the naïve match rate m may serve as a substitute, or flags may be based on m^* or $\widetilde{m}^*$ from higher grades. Note that naïve match rates may also have references from prior years, by comparing $m_{5,2021}$ to $m_{5,2019}$, where the latter uses grade 5 data in 2017 as a denominator. These may also be rescaled to reference 2019 baselines as $\widetilde{m}$.

States should report match rates in noticeable locations in school, district, and state report cards. Users know that this year is unusual and should rightfully be skeptical of any report that does not acknowledge this clearly. A simple statistic that contextualizes and caveats subsequent results is necessary for accurate interpretations.

Metric 2: The Fair Trend

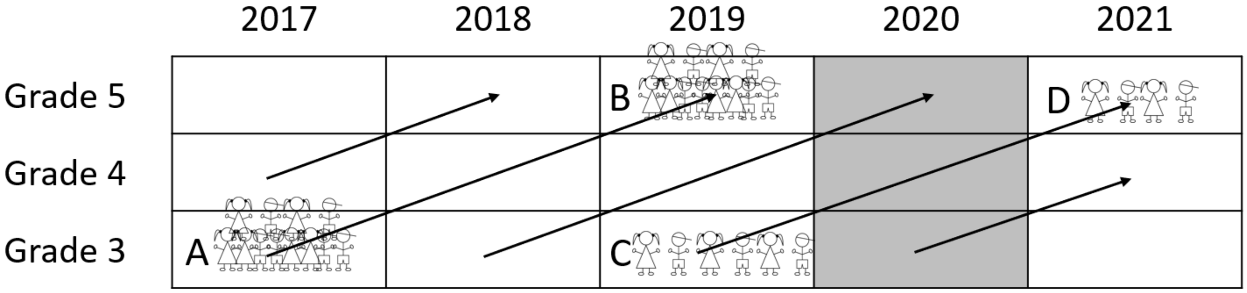
One of the most obvious indicators that schools, districts, and communities need academic support will be substantial declines in their aggregate test scores from recent years. Unfortunately, declines may be caused not by declining academic performance but by large numbers of previously higher scoring students who have left the school system or no longer have comparable scores. This will lead states to target academic support to schools that do not need it. Inversely, stable or rising test scores may be caused not by increases in academic performance but by large numbers of previously lower scoring students who have left the school system or no longer have comparable scores. This will lead states to neglect provisions of academic support to schools that in fact require it.

The second metric, the Fair Trend, enables appropriate comparisons of performance this year to the performance of academic peers two years prior. The Fair Trend includes all observed scores from this year with an appropriate and comparable baseline two years prior. Figure 2 illustrates this using the same cohort diagram as Figure 1. Following this illustration, we can calculate the Fair Trend for grade 5 in 2021 as follows:

1. Identify students in school s and grade 3 in 2019 (Cell C) who have comparable test scores in grade 5 in 2021 (Cell D).[3]
2. Find their same-score "academic peers" in 2017, in grade 3 (Cell A).
3. Report the academic outcomes of these academic peers in 2019 in grade 5 (Cell B). The fair trend compares observed scores in Cell D to the scores of academic peers in Cell B.

[3] Unlike the match rate, which references to past numbers of students in the same school, these scores may be from any school anywhere in the state longitudinal system, not just school s.

Figure 2. *Illustrating the calculation of the "Fair Trend" by comparing scores in 2021 (Cell D) to that of academic peers in Cell B. Academic peers have the same scores in Cell A as current students did in Cell C.*



To operationalize academic peer outcomes in year $y-2$, or 2019, we can fit a flexible model for 2019 scores in terms of 2017 scores. We should fit this model flexibly, nonlinearly, and perhaps nonparametrically. Here, I illustrate the concept simplistically with a linear regression model and leave it to states and their technical advisory committees to select an appropriate model:

$$x_{g,y-2} = \alpha + \beta x_{g-2,y-4} + \varepsilon. \qquad (1)$$

I propose using such a fitted model, whatever the selected functional form, to predict scores for same-score academic peers. There should be a single model for each subject, grade, and year, applied to the entire state population. Define longitudinally matched scores $x^*_{g-2,y-2}$ as the scores from students in the current school, but from two years and two grades prior. Then, use the fitted model parameters from the earlier cohort ($\hat{\alpha}$ and $\hat{\beta}$) to report the academic outcomes for academic peers of these students: $\hat{\alpha} + \hat{\beta} x^*_{g-2,y-2}$. For a "Fair Trend" in terms of average scores, we would compare the following:

Fair average trend:

2021 average score:	$Avg(x_{g,y})$
Academic peer average scores in 2019:	$Avg(\hat{\alpha} + \hat{\beta} x^*_{g-2,y-2})$

For a "Fair Trend" comparison of proficiency rates[4] given stable proficiency cut scores c_g:

Fair proficiency trend:

2021 proficiency rate:	$P(x_{g,y} > c_g)$
Academic peer proficiency rate in 2019:	$P(\hat{\alpha} + \hat{\beta} x^*_{g-2,y-2} > c_g)$

States often report bar charts or line graphs that track proficiency rates of schools, districts, and the state over time. States can use similar models going further back in time to report fair trends over longer time periods.

[4] I do not recommend comparing trends in proficiency rates across groups or schools whose base rates differ (Ho, 2008).

Like the previous match rate metric, there may be some student scores in 2021 that are not matched in the state from 2019, due to in-migration into the state testing system or missed past testing. I designate these scores $x'_{g,y}$ and assume these are relatively few, or at least that they are no more frequent than in any other year. I think these scores should be included even if they are not matched, on the principle of inclusivity.

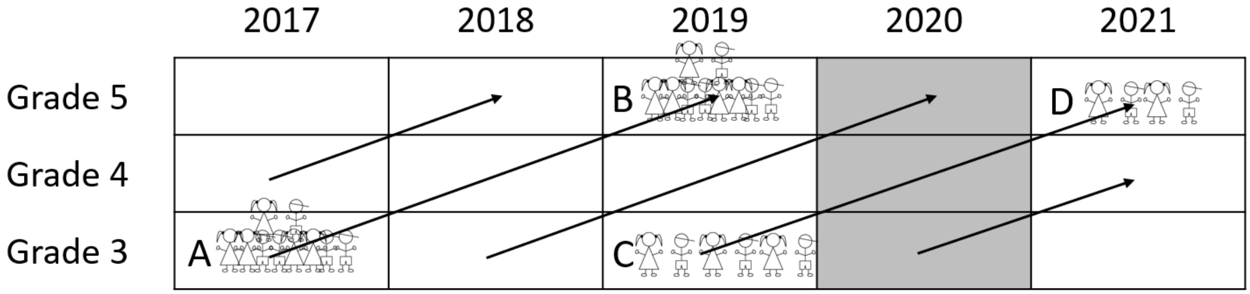
For grades 3 and 4, where no 2-year predictions are possible, various matching algorithms are possible using any array of covariates. In grades 5-8, where we can compare covariate matching and test-score matching directly, we can see whether both approaches produce similar results. If they are similar, we can provide such matching algorithms provisionally. We can also compare discrepancies between the Fair Trend and the potentially biased trend that is observed cross-sectionally, in grades 5-8. If they are similar, states can report potentially biased observed trends in grades 3 and 4. My preference is to foreground "fair trend" results from grades 5-8.

Metric 3: The Equity Check

Perhaps the most important metric from the perspective of documenting potential inequality through this pandemic is the "equity check," the best-case results for students who are not in the system and do not have comparable scores this year. In some ways, this is the complement of the Fair Trend. Instead of comparing current scores to a fair baseline, the Equity Check uses past scores to remind us of the students we are missing. Figure 3 shows the same cohort illustration. I illustrate the Equity Check for 5th graders:

1. Identify students in school s and grade 3 in 2019 (Cell C) who do NOT have comparable test scores in grade 5 in 2021 (Cell D).[5]
2. Find their same-score "academic peers" in 2017, in grade 3 (Cell A).
3. Report the academic outcomes of these academic peers in 2019 in grade 5 (Cell B). These outcomes represent a best-case Equity Check for students in schools in 2019 who no longer have comparable scores.

Figure 3. *Illustrating the calculation of the "Equity Check" as the scores of academic peers in Cell B of students who went missing from Cell C to Cell D. Academic peers have the same scores in Cell A as now-missing students did in Cell C.*



[5] Unlike the match rate, but like the fair trend, these scores may be from any school anywhere in the state longitudinal system, not just school s.

Like the Fair Trend, the Equity Check relies on a flexible model for scores two years prior, estimated from scores four years prior. Define longitudinally *unmatched* scores $x'_{g-2,y-2}$ as the scores from students two years and two grades prior who do not have comparable test scores in the current grade g and year y. Using the same estimated parameters as Equation 1, the best-case Equity Check for 2021 average scores is:

$$Avg\left(\hat{\alpha} + \hat{\beta}x'_{g-2,y-2}\right).$$

And this is the Equity Check in 2021 proficiency rates:

$$P\left(\hat{\alpha} + \hat{\beta}x'_{g-2,y-2} > c_g\right).$$

I recommend comparing Equity Check scores to current year scores to remind users of the scores of students who are missing. I call this a “best case” because the Equity Check assumes academic learning rates for those who went missing from 2019 to 2021 are the same as those in 2017 to 2019. This is obviously optimistic. Thus, I do not consider the Equity Check an estimate of scores for missing students. Instead, I consider the Equity Check as an expression of the past 2019 scores of students we are now missing in 2021, rescaled to the current grade for direct comparison with current-year results.

Some students who are in the Equity Check may have academic peers who were also in missing for reasons unrelated to the pandemic two years ago, due to out-migration from the state public school system or missed testing from 2017 to 2019. We can account for these students roughly by using the same prediction equation, estimating 2019 scores as if they had remained in their 2017 to 2019 cohort. We can then “remove” these academic peers from the Equity Check to account for the fact that some would have departed absent the pandemic. This requires a simple calculation based on weighted averages of academic peer scores. I do not feel this adjustment is necessary. In our current condition, I would rather err on the side of including all students who are in an Equity Check, even those who would have been in an Equity Check under normal conditions.

Like the Fair Trend, the Equity Check is not estimable in grades 3 and 4 without alternative methods for matching. Although I recommend that states explore alternative matching methods to report Equity Checks for these scores, as I recommended in the previous section, my preference is to foreground Equity Check metrics from grades 5-8.

Concluding remarks

This memo reviews simple calculations for three descriptive statistics that I recommend all states report with aggregate report cards in this unprecedented year. The Match Rate is an essential metric for the representativeness of aggregate scores. The Fair Trend is necessary to avoid misclassification of schools that need the most academic support. And the best-case Equity Check reminds us of the academic history of students who would not otherwise be on our radar. Reported carefully, these metrics serve as an academic census, not just the academic status of students we have, but the academic status of students we are missing. With other measures, states can provide a broader educational census that I believe reflects the appropriate role of state education systems in this crisis.

These metrics are not a substitute for standard psychometric research conducted annually to demonstrate the year-to-year comparability of scores. All 3 of the metrics I propose assume that this year's scores are comparable to past years and hold the same meaning in terms of what students know and what students can do academically.

Many states are considering remote administration of tests this spring. This is a serious risk to the comparability of scores. If ongoing or post hoc research finds that states cannot compare scores between in-school and remote modes fairly, the metrics I recommend here will serve as an important hedge to enable accurate interpretations if Match Rates fall. However, even when remote testing rates are low or states conclude scores are comparable, these metrics will be necessary given variable rates of participation and out-migration from schools.

Some states are also considering fall testing. Without previous fall tests to reference as a baseline, fall testing precludes any sensible estimation of Fair Trends. In my opinion, fall testing is not well aligned with the goals of the Department of Education memo.

There are alternative technical approaches to calculating the Match Rate, Fair Trend, and Equity Check, and answering the questions these metrics attempt to answer. In these calculations, I have tried to balance transparency and accuracy. I welcome alternative operationalizations. I encourage states to keep it simple lest risk the public trust on what appears to be a black box. State Technical Advisory Committees and resources like the National Council on Measurement in Education's ongoing free webinar series can provide assistance and perspective. But there is no greater threat to the public trust than reporting statistics like this is business as usual.

States should prepare for these metrics now. States and their vendors can arrange the data in their longitudinal data systems, select an improvement upon the model presented in Equation 1, decide on an approach to early grades 3 and 4, and draft aggregate score reports to include these metrics. Absent these metrics or similar metrics, standard interpretations of aggregate scores will be invalid, and the commendable goal of the Biden administration to use scores to "target resources and supports to the students with the greatest needs" (U.S. Department of Education, 2021, p. 1) will fail.